City Research Online

City, University of London Institutional Repository

Citation: MacFarlane, A. (2007). Evaluation of web search for the information practitioner. Aslib Proceedings; New Information Perspectives, 59(4-5), pp. 352-366. doi: 10.1108/00012530710817573

This is the accepted version of the paper.

This version of the publication may differ from the final published version.

Permanent repository link: https://openaccess.city.ac.uk/id/eprint/4454/

Link to published version: http://dx.doi.org/10.1108/00012530710817573

Copyright: City Research Online aims to make research outputs of City, University of London available to a wider audience. Copyright and Moral Rights remain with the author(s) and/or copyright holders. URLs from City Research Online may be freely distributed and linked to.

Reuse: Copies of full items can be used for personal research or study, educational, or not-for-profit purposes without prior permission or charge. Provided that the authors, title and full bibliographic details are credited, a hyperlink and/or URL is given for the original metadata page and the content is not changed in any way.

City Research Online: http://openaccess.city.ac.uk/ publications@city.ac.uk

Evaluation of Web search for the Information Practitioner

A. MacFarlane
Centre for Interactive Systems Research, Department of Information Science, City University, London, UK
andym@soi.city.ac.uk

Keywords
Web search evaluation precision diagnostic measures

Abstract

Purpose: The aim of the paper is to put forward a structured mechanism for web search evaluation. We point to useful scientific research and show how information practitioners can use these methods in evaluation of search on the web for their users.
Methodology/Approach: The paper puts forward an approach which utilizes traditional laboratory based evaluation measures such as average precision/precision at N documents, augmented with diagnostic measures such as link broken etc which are used to show why precision measures are depressed as well as the quality of the search engines crawling mechanism.
Findings: The paper shows how to use diagnostic measures in conjunction with precision in order to evaluate web search.
Practical implications: The methodology presented in this paper will be useful to any information professional who regularly uses web search as part of their information seeking, and needs to evaluate web search services.
Originality/value: The paper argues that the use of diagnostic measures is essential in web search, as precision measures on their own do not allow a searcher to understand why search results differ between search engines.

1. Introduction

Web search is an important part of the working life of an information professional, but a little understood issue is one of evaluation. How do such professionals evaluate the retrieval effectiveness of a particular information need on a given search engine, or compare search engines. There has been some research in web search evaluation, but few attempts to practically apply evaluation methods in a real environment. There is a need for structured and formal techniques for evaluation that yield quantitative data, in which searchers can clearly see differences in search engines. Such techniques have been around for over 40 years (Aitchison and Cleverdon, 1963) using precision and recall measures, but these techniques do not tackle all the issues which may occur when evaluating web search. In this paper we show why such traditional IR measures on their own do not provide enough information for the researcher when evaluating web search, and to show how diagnostic measures (such as recording the number of broken links) can be used to augment such traditional measures. The paper argues that the methodology put forward gives a much better idea of the retrieval effectiveness of web search engines, as well as the ability to examine other process in web search (such as crawling) which are not part of online search and are not addressed by such measures as recall and precision. The paper is organised as follows. Section 2 describes the previous research done in web search evaluation, which leads on to the motivation in section 3. In section 4 we describe the proposed evaluation methodology, giving an example of an evaluation with the methodology in section 5. A conclusion is given at the end.

2. Previous research in evaluation of Web Search

A great of research has been done on the evaluation of Web Search including various Tracks in the TREC conference series including the VLC2 (Hawking et al, 1999) and Web Tracks (Craswell and Hawking, 2005). Strong arguments are made for the use of scientific methods to evaluate web search either in a live environment (Hawking et al, 2001) or on a static frozen collection (Craswell and Hawking, 2005). A live environment in this context is a real online searching situation where the documents set changes (usually increases in size), while a static frozen collection is a document set used in laboratory style evaluations. Whilst we accept that in order to control variables for scientific experiments, in order to gather useful data, an information professional working with real Web Search needs to work with live web search engines in order to assess the fluid aspects of web search. However the practitioner can still learn from the many valuable lessons from scientific experiments, particularly the measures used by Web IR researchers. But what precisely is it that the practitioner needs to evaluate?

Measure	Calculation	Query Type	References
Average Precision	Average of all precision scores each time a document is retrieved (example of how to calculate this measure is given in table 3).	Informational Transactional	Hawking and Thistlewaite (1998), Hawking et al (1999), Gordon and Pathak (1999), Hawking et al (2000), Hawking et al (2001), Hawking (2001), Hawking & Craswell (2002), Craswell & Hawking (2004), Clarke et al (2004).
Precision at N docs	Divide number of relevant documents by N (where N is the total number of documents retrieved).	Informational Transactional	Hawking and Thistlewaite (1998), Hawking et al (1999), Leighton HV and Srivastava J (1999), Gordon and Pathak (1999), Wu and Li (1999), Hawking et al (2000), Hawking et al (2001), Hawking (2001), Hawking & Craswell (2002), Hawking & Craswell (2003), Hawking et al (2004), Craswell & Hawking (2004).
Mean Reciprocal Rank	Decending scale from (say) 1-5 is used, e.g hit at rank 1 is given score 1.0, at 2 score 0.5 etc. hits outside of rank 5 are assigned 0 score. Scores are then averaged from all queries.	Navigational	Chowdhury and Soboroff (2002), Hawking & Craswell (2002), Hawking & Craswell (2003), Hawking et al (2004), Craswell & Hawking (2004).
R-Precision	Precision at the R (total number of relevant documents)	Informational Transactional	Hawking et al (2004).
Success N	Proportion of queries in which good answers were found at rank N	Informational Transactional Navigational	Craswell & Hawking (2005).
Recall N	Recall at N documents retrieved	Informational Transactional	Craswell & Hawking (2005).
% top N	Proportion of queries where the right answer was found in the top N hits	Navigational	Hawking & Craswell (2002), Hawking et al (2004).
% fail N	Proportion of queries where no right answers was found in the top N hits	Navigational	Hawking & Craswell (2002).

Table 1 – Traditional IR Measures used to evaluate search

Web queries can be divided up into three main types: navigational, transactional and informational (Broder, 2002). A navigational query is one in which

a user wants to find a particular web site (e.g. the home page of City University), whereas a transactional query is where the user wants to find a site where some further interaction will take place (e.g. where can I buy bookcases). An informational query is one which is needed to satisfy an Anomalous State of Knowledge or ASK (Belkin et al, 1982). An example of this would be "what are the legal precedents for civil cases in conveyancing?". With navigational queries the user is looking for a single item for the most part, while the user requires multiple items for transactional and informational queries. The practitioner could potentially be faced with both navigational and transactional queries, but by in large their users information needs will require informational queries. Table 1 describes some traditional IR measures used for evaluation in Web Search together with their target query type. It should be noted that binary decisions on relevance (relevant or not-relevant) are dominant in the field to date, however there is some interest in using non-binary evaluation methods for web search (Clarke et al, 2004).

Other non-traditional methods have been used for evaluation purposes e.g. Vaughan (2004) uses a number of different mechanisms for example result ranking quality (e.g. correlation between user ranked pages and pages ranked automatically by search engines), and stability measures (compare the ranking of documents over a given period e.g. two weeks). However these measures rely for the most part on relevant documents (in effect a detailed comparison of precision) or the ranking mechanism (between two sets of results). Other more frequently used measures try to examine in more detail why particular documents are not relevant, and if precision is effected adversely – we label these 'Diagnostic measures', see table 2.

Measure	Calculation	Query Type	References
Duplicates	Count the number of duplicate documents in the top N hits	Informational Transactional	Leighton HV and Srivastava J (1999), Wu and Li (1999), Oppenheim et al (2000)
Broken Links	Count the number of broken Links (e.g. 404 not found)	Informational Transactional Navigational	Wu and Li (1999), Oppenheim et al (2000)

Table 2 – Diagnostic measures

These methods have been used in difference contexts e.g. Leighton and Srivastava (1999) look at Web Search generally while Wu and Li (1999) focus on Web Search for Health Information. It should be noted that these Diagnostic measures can be used in conjunction with non-binary relevance judgements. In a review of web search methods Oppenheim et al (2000) argue for a broad based approach using both traditional and diagnostic measures. The author fully agrees with this strategy, but if you are to use Diagnostic measures you need to consider other aspects of web search e.g. none of the studies referenced in this paper consider Spam documents.

3. Motivation for the study

The primary motivation for this study is to give information practitioners or professional search intermediaries some guidance on how to evaluate Web Search in the light of experience gained by researchers in the field. We believe that many of the traditional IR measures which have been used for many years are still useful for evaluation in our context, but many of the studies concentrate on evaluating test collections. This is very useful in a scientific context, but information practitioners have to deal with real dynamic collections, and these traditional measures in isolation

do not provide the required mechanism for dealing with Web Search. The Diagnostic measures provide information practitioners with the extra data the need in order to properly evaluate searching for information on the web for their users. The rest of the paper outlines this evaluation methodology and how to use it.

4. Proposed evaluation methodology

Before starting the evaluation the practitioner needs to make some important decisions on what they will be evaluating. The first (and most obvious) decision to make is to determine which search engines they will evaluate in their study. This may include mainstream search engines such as Google and Altavista or more specialist search engines such as the Health Library and Law Crawler. The choice will be often determined by the information needs of the users the practitioner serves. The number of search engines to evaluate is also an important issue – this will depend on the resources available to the practitioner (the evaluation methodology described in this paper is a time consuming process).

When this is decided the number of topics to evaluate needs to be chosen (we suggest 50 as used in many TREC experiments) and the number of pages to examine for each topic for every search engine (this should be consistent across all topics and search engines). For the latter we recommend that only the first top ten sites are examined in the evaluation as it reduces the evaluation workload significantly. Many case studies have shown that users very rarely view pages of hits beyond the first page of the hitlist e.g. in Silverstein et al (1999) 85.2% of individual queries only viewed one page of results from AltaVista.

A further issue to consider is the number of URL's to navigate from a hit list to find a relevant web page – or alternatively the number of clicks the user requires to find that web page. This may be needed for some queries where a relevant web page, satisfying the users information need is buried somewhere within the web site. A maximum of three clicks to find a relevant web page is considered reasonable. Alternatively the practitioner can use a more stringent method and assume that the user is only interested in pages that are linked directly from the hit list. In all cases the strategy used should be consistent, to ensure that the results produced for the evaluation make sense.

When the issue of what to evaluate has been decided on, the practitioner can then think about conducting the evaluation. In the next three sections we describe the measures and the process that can be used for Web Search evaluation. Note that for the rest of our discussion we make a binary relevance assumption for documents.

4.1 Traditional Measures

We recommend the use of two measures: Precision at N documents retrieved and average precision, as described in section 2. We used the Precision at N calculation to see how precision deteriorates over a given number of blocks or chunks. If the recommendation on examining only the first ten hits is taken from above, a reasonable strategy is to calculate precision at 5 and 10 sites retrieved. These are standard measures used for many years in laboratory based evaluations, and have been used on all kinds of information including news stories, government report, journal articles as well as the web. Calculation of precision at 5 and 10 is very simple, the total number of relevant documents found to that point is divided by either 5 or 10.

How well does the engine retrieve documents against the known total number of documents relevant (recall)? It is impossible to know the recall for collections the size of the Web, so we need some estimate that we can sensibly use in order to give a

figure to compare. If practitioner inspects at ten documents at most, they can make the assumption that we have at least 10 documents for our given information need. This strategy might come in for some criticism in the sense that how can you be sure that there are 10 relevant documents for any information need you may have? The authors answer to this is that there are now 8 billion odd web pages indexed by Google as this paper is being written, and it is reasonable to assume that at least 10 of those will be relevant for many users information needs. If the practitioner is unhappy with this mechanism, they could use a pooling method for relevant documents (Voorhees and Harman, 2000), which would mean that the retrieved sets would need to be merged and the number of relevant documents found for each topic used instead of the assumed 10. This does however place an extra burden on the practitioner when conducting the evaluation. The author does not regard this as a significant issue provided the strategy taken on the assumption of relevant documents is consistent. Because of the assumption made with regard to relevant documents we label the measure 'Estimated Average Precision'.

We use this assumption to calculate average precision (see table 1 above). Average precision is a precision-based measure linked to recall. The evaluator uses this measure to see how our search engines are doing against the estimated recall and how this relates to precision. It also tells the evaluator how well relevant documents are being ranked across the whole hit list. Table 3 shows how average precision can be calculated (given our assumptions on 10 documents retrieved, 10 documents relevant):

RANK	RELEVANT	RELS/RANK
1	1	1/1 = 1
2	0	-
3	1	2/3 = 0.67
4	0	-
5	1	3/5=0.6
6	0	-
7	0	-
8	1	4/8=0.5
9	1	5/9=0.56
10	0	-

Table 3 – Calculating Estimated Average Precision

Each time a relevant document is retrieved, the total number of relevant documents found so far are divided by the current rank. The evaluator then accumulates the average precision scores which in the case of table 3 gives us a total of 3.33. Dividing this by ten (our assumed number of relevant documents) gives us an Estimated Average Precision (EAP) of 0.33. The more relevant documents higher up the hit list rank, the better the EAP score.

4.2 Diagnostic measures

These measures are used to show why documents are not relevant beyond the fact that many documents do not meet an information need, and the subsequent impact this has on the precision measures described in section 4.1. Two diagnostic measures have already been introduced, Duplicates and Broken Links, but there are other measures which need to be considered such as Spam and a figure for hit lists which do not retrieve a full 10 documents (which we must consider if we assume that there are at least 10 relevant documents). The calculation for these measures are simple – the

occurrences of a particular metric is accumulated (scores are recorded between 0 and 10). We describe each of these diagnostic measures in turn below.

Repeated Documents (or duplicates)
It is often the case that searches will bring up identical pages in a retrieved list. Since they contain the same information, it makes sense to mark the first encountered page relevant, and treat other subsequent page as being irrelevant. Choosing criteria for duplicates can be difficult – must the documents be identical in every sense, is the information in the document identical, are retrieved documents from the same site (when you only actually want one when completing navigational searches)? A good example of why this may happen is multi-national companies that have offices in several places – some search engines are better at handling this type of problem than others. It may be best to use a simple method – if the page looks the same and has the same information it's a duplicate, otherwise it's not. However the definition of duplicates will often depend on the type of information being searched for and the query type.

Not retrieved
As we want 10 documents to retrieve, any hit lists which retrieve less then 10 documents damages our precision and we want to penalise search engines that do not retrieve our required number of pages.

Link broken
This occurs when a user clicks on a link and you get an error message e.g. 404 not found. Sometimes you may find that the link returned is a redirected page – the author would suggest that if the target of the redirection is relevant, then you mark the page as being relevant (if the webmaster/author has taken the trouble to make sure the information is available we should give them credit). In such a case the evaluator should not mark the link being broken.

Spam
A big issue for search engines is Web page designers putting in words that bare no relation to the content of a page. This can be done in the Meta tags in HTML or by putting the words in the main body of the document, but using a font/colour that makes it invisible on the browser. This means that when a user types in their search words, they retrieve documents/pages that are completely irrelevant to their information needs. The user is puzzled, as it is obvious that the page is irrelevant, and they cannot find any trace of their search words in the retrieved page. These pages are called "Spam" pages and they can be very annoying to the user. This technique tends to be used by the 'Adult Entertainment' industry and there is something of an arms race between web search engines and such organisations. Spam pages harm precision of course (they are not relevant) so should be recorded. A good survey of Spamming techniques can be found in Henzinger et al (2002).

4.3 The process of evaluating Web Search
The simple evaluation procedure for this type of experiment is as follows:

- Use a given query on all the search engines.
- Judge each engine for this query, and record the results of each measure.

- When all the results for all the queries applied to all the search engines calculate the average for every search engine on every measure.
- Tabulate each measure separately, listing the search engine and its score on that measure.
- Apply statistical techniques to find significant differences between the effectiveness of search engines.
- Compare and contrast each search engine for each measure to see how well search engines did against each other: using the diagnostic measures to show why precision was reduced for any search engine.

5. An Evaluation experiment

5.1 Data used for the experiment

We conducted an evaluation of 50 queries whist working for a commercial organisation in 2000 (the queries are declared in Appendix 1). The queries are mostly taken from the logs of a now defunct Web search engine; the author added some informational queries to the set. We used the same method for choosing queries from the log as used at the TREC-8 Web Track (Hawking et al, 2000): that is we inspected a number of queries and picked those which we felt confident that we understood what the user was searching for and could therefore make appropriate relevance assessments. The average number of terms for the query set used is 2.68: this is about what you would expect from a set of web queries and is not far off the figure quoted by both Silverstein et al (1999) and Jansen et al (1998) of 2.35 terms per query. Our classification of the query set found that 18% were navigational, 46% transactional and 36% informational. This is a reasonable distribution of the queries for our experiment, as there are enough web type searches to be close to the type of searches that most web users will undertake (64% for navigational/transactional queries). However, there are a sufficient number of informational queries to make the study of interest to practitioners whose users are more likely to require the resolution of information needs.

We used all the assumptions and techniques for this evaluation declared and described in section 4. We did not inspect the URL beyond the first click: our requirement was the URL's in the hit list should contain relevant information (Hub sites were not therefore considered).

5.2 Experimental Results

The Precision results collected for the experiment are declared in Table 4. We use these results to show how the measures can be used in practice.

Search Engine	P@5	P@10	Average Precision	Spam	Dups	Link Broken	Not Retrieved
Google	0.424	0.386	0.290	0	0.82	0.50	0.20
AltaVista	0.280	0.256	0.178	0	0.28	0.72	0.00
Lycos	0.184	0.160	0.093	0	2.02	1.12	0.54
Yahoo	0.318	0.280	0.190	0	0.44	0.34	2.22

Table 4 – Evaluation Results

What stands out in the results is that no Spam documents were retrieved, and we can therefore discount Spam as a problem for this particular set of queries for a given time period. This could be because many of the queries are quite esoteric or it could be that

search engines were doing a good job of detecting Spam at that time. Google clearly comes out on top with respect to all the precision measures, and quite clearly did a lot better than the other search engines for ranking documents e.g. Google provided a third better precision at 5 documents retrieved than its nearest rival Yahoo.

The worst performer on this set of queries using the precision measures is Lycos, and it is clear why this is the case from evidence provided by the diagnostic measures: on average 2 hits were duplicates while more than one broken link per query was found. This clearly demonstrates the value of diagnostic measures and the impact they have on precision. However some diagnostic measures are useful to examine other aspects of web search e.g. the Link Broken measure demonstrates that Lycos web crawling mechanism was not as effective as the others in 2000. Yahoo recorded the worst 'Not Retrieved' score and did not do as well as Google largely because it was not retrieving the required 10 documents (missing over 2 sites per query on average). Interestingly only Lycos did worse on the duplicates measure than Google, and Yahoo has the least number of broken links of all the search engines. Overall the conclusion is for this set of queries for that particular time period, Google was the most effective search engine and did not have as much of a problem as the other search engines with respect to diagnostic measures. Lycos overall is the search engine which has its precision results most adversely effected by errors recorded by diagnostic measures. We examine these figures in terms of statistical significance below.

5.3 Significance testing on Precision Results

Measure	**Google vs. AltaVista**		**Google vs. Yahoo**		**Google vs. Lycos**		**AltaVista vs. Yahoo**		**AltaVista vs. Lycos**		**Yahoo vs. Lycos**	
	t-test	wlc	t-test	wlc	t-test	wlc	t-test	wlc	t-test	wlc	t-test	wlc
P@5	.001	.002	.015	.024	.000	.000	**.472**	**.567**	.021	.033	.005	.024
P@10	.000	.000	.001	.000	.000	.000	**.573**	**.414**	.004	.006	.003	.007
Ave Prec	.000	.001	.000	.000	.000	.000	**.759**	**.413**	.004	.025	.005	.004

Table 5 – Significance tests on Precision results
(figures in **Bold** are <u>not</u> statistically significant)

It should be noted that an increase in precision does not necessarily mean that there is a real difference between search engines i.e. one web search engine is shown to provide better retrieval effectiveness. In order to do this some kind of significance testing is useful, but there is some controversy on this issue. Some argue (van Rijsbergen, 1979) that parametric tests such as the t-test are not applicable as the form of the underlying distribution (of relevant documents) is unknown. Others such as Hull (1993) and Sanderson & Zobel (2005) argue that parametric measures such as the t-test can be used even if the assumption on the underlying data having a normal distribution is violated. One method around this is to use a non-parametric test such as the Wilcoxon test (Hull, 1993) in conjunction with the t-test and only accept that there is a difference between the two systems if both measures agree that the difference is significant.

This is the method we used on the data collected in the experiments (see table 5). T-test results are marked as 't-test' while Wilcoxon test results are marked as 'wlc' in the table. The practitioner does not need to know the details of these tests, just that a result below 0.05 is regarded as being significant, while a result of 0.01 or below can be regarded as highly significant (Rowntree, 1981). Many such statistical tests are

available in Microsoft's Excel spreadsheet software or can be downloaded from the web. It can be seen from table 5 that both tests agree on what is significant and what is not significant, which gives us a little more confidence on any conclusions we draw from this data. Given this we can see that both tests are in agreement that Google provides a retrieval effectiveness improvement over the other search engines which is highly significant for the most part; apart from the test against Yahoo on 5 documents retrieved. These tests give us more confidence that the retrieval effectiveness Google provided over the other search engines used is actually real (in section 5.2 above). It should be noted that both tests do not agree on what differences are significant and what are highly significant: e.g. Yahoo vs. Lycos at 5 documents retrieved. In cases where the measures do not agree we recommend that the practitioner err on the side of caution when drawing any conclusions.

Measure	**Google vs. AltaVista**	**Google vs. Yahoo**	**Google vs. Lycos**	**AltaVista vs. Yahoo**	**AltaVista vs. Lycos**	**Yahoo vs. Lycos**
P@5	51.4	33.3	130.4	13.6	52.2	72.8
P@10	50.8	37.9	141.3	9.4	60.0	75.0
Ave Prec	63.1	52.4	212.9	7.0	91.9	105.4

Table 6 – Percentage improvement for best result: Precision results

Practitioners should be wary of using percentage increases in precision to differentiate between search engines (Sanderson & Zobel, 2005). A good example of this can be found in Table 6. It can seem that many of the increases in precision (particularly for Google over the other search engines) are very impressive. The percentage increase from AltaVista to Yahoo is also quite good (7% for average precision). However using the data from the significance tests applied, any difference between AltaVista and Yahoo is not regarded as being significant, even though on the surface Yahoo would appear to be the better search engine for the query set used.

5.4 Significance testing on Diagnostic measures

Table 7 declares the results of significance tests on the Diagnostic measures from table 4. As with precision measures, there is complete agreement between the two statistical tests as to which pairwise comparisons are significant. This is very encouraging indeed and gives us yet more confidence on any statements we may make on statistical significance, with respect to all the measures. The measure also distinguishes between most of the comparisons between significant and highly significant differences, apart from the Link Broken measure on Google vs. Lycos and AltaVista vs. Yahoo.

Measure	**Google vs. AltaVista**		**Google vs. Yahoo**		**Google vs. Lycos**		**AltaVista vs. Yahoo**		**AltaVista vs. Lycos**		**Yahoo vs. Lycos**	
	t-test	wlc	t-test	wlc	t-test	wlc	t-test	wlc	t-test	wlc	t-test	wlc
Dups	.003	.002	.010	.020	.002	.005	**.344**	**.294**	.000	.000	.000	.000
Link Broken	**.132**	**.167**	**.242**	**.353**	.011	.009	.008	.013	**.058**	**.074**	.001	.001
Not Ret.	**.322**	**1.00**	.001	.000	**.101**	**.250**	.000	.000	**.060**	**.125**	.003	.002

Table 7 – Significance tests on Diagnostic results
(figures in **Bold** are not statistically significant)

With respect to the statements made in section 5.2 with regard to diagnostic measures and their impact on precision, it is clear that for the most part that the differences appear to be statistically significant for most of the worst performing search engines when completing pairwise comparisons. Yahoo's 'Not Retrieved' and Lycos's 'Duplicates' figures when compared to other search engines results can be regarded as being very significantly different. The statement made about the reason for Yahoo's depressed precision against Google because of the 'Not Retrieved' measure is validated by these test results. However when comparing Lycos against AltaVista on both the 'Link Broken' and 'Not Retrieved' measures, we find that there is no statistical evidence of difference between the two search engines. The statistical significance recorded on the precision measures between these two engines must therefore by down largely to the poor performance of Lycos on the 'Duplicates' measure. We can draw a few other conclusions about differences in retrieval effectiveness from many other pairwise comparisons, which allow us to show which diagnostic measures are most likely to have an effect on precision e.g. with Google and AltaVista, 'Duplicates' appears to be the most likely reason between the difference in retrieval effectiveness. Only on one occasion (Yahoo vs. Lycos), do all three diagnostic measures appear to have an effect.

Measure	Google vs. AltaVista	Google vs. Yahoo	Google vs. Lycos	AltaVista vs. Yahoo	AltaVista vs. Lycos	Yahoo vs. Lycos
Dups	65.9	86.4	146.3	57.1	621.4	359.1
Link Broken	30.6	47.1	124.0	111.8	55.6	229.4
Not Ret.	*Inf*	1,010.0	170.0	*Inf*	*Inf*	311.1

Table 8 – Percentage improvement for best result: Diagnostic results

It can be seen from Table 8 that using percentage improvements is a completely inappropriate method for distinguishing between search engine performance on diagnostic measures. A good example of this is the result from AltaVista on the 'Not Retrieved' measure: as this was zero any comparison between AltaVista and other search engines on this measure is rendered meaningless. Further evidence (if needed) is provided by the comparison between Google and Lycos on the 'Not Retrieved' measure: an increase of 170% is recorded from Google to Lycos, but both the t-test and Wilcoxon measures agree that the difference is not statistically significant. One of the main reasons for this behaviour is that diagnostic measures are not normalised like precision measures (between 0-1) and are therefore more sensitive to any increase.

6. Conclusion

The evaluation methodology presented in this paper is a practical (if labour intensive) mechanism for evaluation, which has been successfully used for teaching purposes at City University for the past four years. The source of this methodology was the need of a commercial organisation, which required an evaluation of search engine technology – this inspired the author to develop the methodology. The author found the method very useful when he applied it and Information Science students at City University London have had the same experience in their working environments, having learnt the method in their information retrieval module. We therefore believe that information practitioners will find this method a useful way of evaluating web search engines for the searches they conduct on behalf of their users. The advantage of this methodology is that it builds on a significant amount of work by the academic

community, and it gives the evaluator much more information on why search engines do not do so well on average using evidence provided by the diagnostic measures. The example evaluation in section 5 demonstrates this clearly, where the impact of diagnostic measures on precision is shown to be significant in many cases. Further work from this study would include measuring the direct impact of diagnostic results on precision for a single search engine using some form of statistical analysis (as apposed to the pairwise comparison method used in this paper).

Acknowledgements

The author is grateful to Prof. Stephen Robertson for advice on both what measures to use and how to interpret statistical significance on the experiments described in this paper.

References

Aitchison, J., and Cleverdon, C. (1963). "A report on a test of the index of metallurgical literature of Western Reserve University". The College of Aeronautics, Cranfield, UK.

Belkin, N., Oddy, R., and Brooks, H. (1982). ASK for Information Retrieval: Part 1. Background and Theory. Journal of Documentation 38(2) [reprinted in Spark-Jones, K. and Willett, P. (1997). Readings in Information Retrieval, Morgan Kaufmann, 299-304]

Broder, A. (2002). A taxonomy of Web Search, SIGIR Forum, Fall 2002, 36(2), 3-10.

Chowdhury, A, and Soboroff, I. (2002). Automatic Evaluation of World Wide Web Search Services, In: Beaulieu, M., Baeza-Yates, R., Myaeng, S and Jarvelin K., (eds), Proceedings of the Twenty-Fifth Annual International ACM SIGIR Conference on Research and Development in Infomration Retrieval (SIGIR 2002), 421-422.

Clarke, C., Craswell, N. and Soboroff, I. (2005). Overview of the TREC 2004 Terabyte Track. In: Voorhees, E. and Buckland L (eds) NIST Special Publication 500-261: The Eleventh Text REtrieval Conference, (to appear).

Craswell, N. and Hawking, D. (2005). Overview of the TREC 2004 Web Track. In: Voorhees, E. and Buckland L (eds) NIST Special Publication 500-261: The Eleventh Text REtrieval Conference, (to appear).

Gordon M and Pathak P (1999) Finding information on the world wide web: The retrieval effectiveness of search engines. Information Processing and Management 35(2) 141-180.

Hawking, D., (2001). Overview of the TREC-9 Web Track: In: (eds). In: Voorhees, E. and Harman D (eds) NIST Special Publication 500-249: The Ninth Text REtrieval Conference, 87-103.

Hawking, D. and Craswell, N. (2002). Overview of the TREC 2001 Web Track. In: Voorhees, E. and Harman D (eds) NIST Special Publication 500-250: The Tenth Text REtrieval Conference, 61-67.

Hawking, D. and Craswell, N. (2003). Overview of the TREC 2002 Web Track. In: Voorhees, E. and Buckland L (eds) NIST Special Publication 500-251: The Eleventh Text REtrieval Conference.

Hawking, D. Craswell, N., and Thistlewaite, P (1999). Overview of the TREC-7 Very Large Collection Track. In: Voorhees, E. and Harman D (eds) NIST Special Publication 500-242: The Seventh Text REtrieval Conference, 91-104.

Hawking, D., Craswell, N., Bailey, P., and Griffiths, K. (2001). Measuring Search Engine Quality, Information Retrieval, 4 33-59.

Hawking, D. Craswell, N., Wilkinson R and Wu, M. (2004). Overview of the TREC 2003 Web Track. In: Voorhees, E. and Buckland L (eds) NIST Special Publication 500-255: The Twelth Text REtrieval Conference, 78-94.

Hawking and Thistlewaite (1998) Overview of the TREC-6 Very Large Collection Track. In: Voorhees, E. and Harman D (eds) NIST Special Publication 500-242: The Sixth Text REtrieval Conference, 93-106.

Hawking, D., Voorhees, E., and Craswell, N., (2000). Overview of the TREC-8 Web Track. In: Voorhees, E. and Harman D (eds) NIST Special Publication 500-246: The Eighth Text REtrieval Conference 131-150.

Henzinger, M, Motwani, R. and Silverstein C. (2002). Challenges in Web Search Engines, SIGIR Forum, Fall 2002, 36(2), 11-22.

Hull, D. (1993). Using statistical testing in the evaluation of retrieval experiments. In: Korfage, R, Rasmussen, E. and Willett, P. (eds). Proccedings of the 16th Annual ACM conference on Research and Development in Information Retrieval, SIGIR'93, 329-338.

Jansen, B, Spink, A, Bateman, J., and Saracevic, T. (1998). Real life information retrieval: A study of User Queries on the web, SIGIR Forum, Spring 1998, 32(1), 5-17.

Leighton HV and Srivastava J (1999) First 20 precision among world wide web search services (search engines). Journal of the American Society for Information Science 50(10), 870-881 .

Oppenheim, C., Morris, A., McKnight, C. and Lowley, S., (2000). The Evaluation of WWW Search Engines, Journal of Documentation , 56(2), 190-211.

Rowntree, D. (1981). Statistics without tears: An introduction for non-mathemeticians. Penguin Books.

Sanderson, M. and Zobel, J. (2005). Information retrieval systems evalation: effort, senstivity, and reliabiity. In: (eds). Proccedings of the 28th Annual International ACM conference on Research and Development in Information Retrieval, SIGIR 2005 162-169.

Silverstein, S., Henzinger, M., Marais, H. and Moricz (1999). Analysis of a very large web search engine query log, SIGIR Forum, Fall 1999, 33(1), 6-12.

Van Rijsbergen, C. (1979). Information Retrieval, 2nd Edition, Butterworths [available on: http://www.dcs.gla.ac.uk/Keith/Preface.html: visited 2 November 2005]

Vaughan, L (2004). New measurements for search engine evaluation proposed and tested, Information Processing and Management, 40(4), 677-691.

Voorhees, E. and Harman D (2000). Overview of the Eighth Text REtrieval Conference. In: Voorhees, E. and Harman D (eds) NIST Special Publication 500-246: The Eighth Text REtrieval Conference, 1-24.

Wu, G. and Li, J. (1999). Comparing Web search performance in searching consumer health information: evaluation and recommendations, Bulletin of the Medical Library Association 87(4) 456-461.

Web References

AltaVista: www.altavista.com [Visited 20th September 2005]

Google: www.google.com [Visited 20th September 2005]

Law Crawler: lawcrawler.findlaw.com [Visited 8th September 2005]

SIGIR Forum Online: www.sigir.org/forum [Visited 8th September 2005]

The Health Library: www.health-library.com [Visited 8th September 2005]

Lycos: www.lycos.com [Visited 20th September 2005]

Yahoo: www.yahoo.com [Visited 20th September 2005]

Appendix 1 – List of Queries used for Evaluation

1. sade adu biography
2. middle east crisis
3. parallel computing
4. information retrieval
5. karl popper
6. philosophy science
7. scramble africa
8. origins second world war
9. urbanwear streetwear urbanclothing hiphop clothing
10. flower arranging
11. mountain climbing safety equipment
12. loft insulation
13. body building
14. arvo part compositions
15. norman conquest
16. meiji restoration japan
17. atomic clock accuracy
18. curry
19. led zeppelin
20. levi jeans
21. bookcase suppliers
22. tour operators spain
23. fiction novel the silver city bombay street children
24. lou reed interview
25. soprano singing
26. gene therapy
27. martin scorsese
28. submersible pump manufacturer germany
29. door fittings
30. currency conversion
31. serbian mafia
32. investments software
33. bodyguard training tuition
34. pictures Linda Lusardi
35. research forest pathology
36. bonsai styles world
37. social security rates
38. land rover defender
39. air fares germany britain
40. telemetry alarm system
41. nokia phone
42. le surete french police
43. restaurants kids central london
44. autonomy
45. microwave ovens
46. engineer jobs uk
47. soorento italy images
48. festival diwali
49. woodpigeon shooting
50. sex